2017

Diagnostic Assessment of L2 Chinese Learners' Reading Comprehension Ability

Shuai Li
Georgia State University, sli12@gsu.edu

Jin Wang
Kennesaw State University, jwang@kennesaw.edu

Li, S., & Wang J. (2017). Diagnostic assessment of L2 Chinese learners' reading comprehension ability. In Zhang, D. & Lin, C. (Eds.). *Chinese as a second language assessment* (pp. 183–202). Springer.

Chapter 9. Diagnostic Assessment of L2 Chinese Learners' Reading Comprehension Ability

Shuai Li
Georgia State University
sli12@gsu.edu

Jing Wang
Kennesaw State University

ABSTRACT

This study explored the feasibility of applying the Rule Space Method (RSM) to diagnosing the strengths and weaknesses of reading ability among learners of Chinese based on their performance on the reading comprehension section of a standardized Chinese proficiency test, the C. Test. Combining literature review, instructor coding, and expert judgment, we finalized a set of eight attributes measured by 30 multiple-choice reading comprehension test items. Eight hundred and fifty seven (857) examinees took the above mentioned test, and their responses to the 30 test items were used for statistical analyses. The results showed that 90.54% of the examinees were successfully classified into one of the pre-specified attribute-mastery patterns, based on which we were able to offer detailed diagnostic reports to individual examinees regarding their mastery/non-mastery of the attributes.

Keywords: Diagnostic language assessment, rule space method, L2 Chinese, reading ability

INTRODUCTION

Diagnostic language assessment (DLA), understood as the "processes of identifying test-takers' (or learners') weakness, as well as their strengths, in a targeted domain of linguistic and communicative competence and providing specific diagnostic feedback and (guidance for) remedial learning" (Lee, 2015, p. 5), has attracted a lot of attention in applied linguistics. For example, the 2015 special issue of *Language Testing* and the 2009 special issue of *Language Assessment Quarterly* were devoted to understanding the various approaches to DLA and their applications to second language (L2) assessment. The surge of interest and empirical effort in DLA is in response to the growing demand from practitioners and stakeholders of language teaching and learning calling for refined assessment techniques that are able to provide individualized diagnoses of test takers' mastery and non-mastery of knowledge and skills in order to guide subsequent teaching and learning (Jang, 2009a; Kim, 2015; Lee, 2015). In this regard, traditional language assessment techniques, such as those informed by classical testing

theories, which typically report examinees' standardized total test scores and section scores (e.g., for reading and listening), are not able to meet this demand. In fact, it is possible that two test takers with different underlying skill/knowledge profiles receive identical total/section scores based on their test performance (Tatsuoka, 2009). Hence, unless there is a means to detect the mastery/non-mastery of latent knowledge/skills, we are not able to conduct individualized remedial teaching and learning for test takers.

The Rule Space Method (RSM) (Tatsuoka, 1983, 1995, 2009), a psychometrically-based technique for tapping latent cognitive *attributes* (defined as knowledge and cognitive processing skills), provides a viable solution to the aforementioned problem. As a statistical method of pattern recognition and classification, the RSM aims to classify examinees' observable test item response patterns into a set of predetermined attribute mastery/non-mastery patterns, called *knowledge states*. In so doing, it can provide fine-grained diagnostic information for individual test takers regarding their strengths and weaknesses in the knowledge and cognitive skills assessed by a test. In the field of second language assessment, the RSM and related methods (e.g., the Fusion Model) have been used to diagnose the knowledge states of examinees as they respond to test items assessing listening and reading comprehension (e.g., Buck & Tatsuoka, 1998; Buck, Tatsuoka, & Kostin, 1997; Jang, 2009a, 2009b; Kim, 2015). Previous studies have mostly relied on existing tests (e.g., TOEIC, LanguEdge, TOEFL iBT), and it is interesting that among those studies targeting the same language skill (e.g., reading), the attributes identified and the knowledge states examined were often different and dependent on the particular test items under investigation. Research is thus needed to examine additional tests to evaluate the generalizability of previous research findings. A related issue is that, because previous studies have exclusively focused on English as the target language, it is critical to expand this line of research to other, particularly those typologically different, languages such as Chinese.

This study is an effort in this direction. It explored the feasibility of using the RSM for conducting diagnostic assessment of test takers' strengths and weaknesses in reading ability as they responded to a standardized Chinese proficiency test, the C. Test. The following sections will first introduce the rationale and procedures of the RSM, followed by a discussion of the applications of the RSM to L2 assessment.

Rule Space Method (RSM): Rationale and Procedures

The Rule Space Method (RSM) (Tatsuoka, 1983, 1995, 2009) was developed with the purpose of reporting fine-grained information about an individual examinee's mastery/non-mastery of specified latent *attributes* (i.e., knowledge and cognitive skills) based on his/her performance on a set of test items. The rationale is that a correct (or incorrect) response to a test item entails the mastery (or non-mastery) of certain latent attribute(s). Therefore, a specific test item can be described by the latent attribute(s) that it measures, and a specific set of test items can be described by different combinations (or patterns) of latent attributes that they measure. Hence,

ideally, by examining an individual test taker's observable item response patterns, one can identify his/her (unobservable) attribute mastery pattern (i.e., *knowledge state*) by means of pattern recognition. In reality, however, test takers' performance on test items are influenced not just by those specified latent attributes, but also by many other factors (e.g., carelessness). Therefore, the RSM also involves a pattern classification procedure which is probability-based. In other words, as Tatsuoka (2009) summarizes, "RSM converts students' item response patterns into attribute mastery probabilities" (p. xii).

The application of RSM involves three phases (Buck & Tatsuoka, 1998; Buck et al., 1997; Gierl, 2007): (1) identifying attributes and determining ideal knowledge states; (2) formulating a classification space (or *rule space*); and (3) classifying examinee responses. During the first phase, test items are analyzed to identify the attributes that need to be mastered for correct responses[1]. This analysis typically involves domain experts' evaluation of test items based on relevant theories and empirical results, occasionally supplemented by an examination of test takers' verbal protocols (e.g., Jang, 2009b). The hierarchical relations (if any) among the identified attributes are then described. For example, Figure 1 illustrates the hierarchical structure of a hypothetical set of five attributes assessed by a collection of test items. As can be seen, the mastery of attribute A1 serves as the prerequisite for the mastery of attribute A2; the mastery of attribute A2, in turn, is the prerequisite for the mastery of attribute A4.

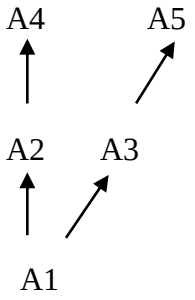

Figure 1. Hierarchical structure of five hypothetical attributes assessed by a test.

With the above information, we can construct an *adjacency matrix* (A) where all (unidirectional) direct relations among the attributes are represented by "1" and the lack of such relation by "0" (Table 1). Through *Boolean* addition and multiplication based on the A matrix (Tatsuoka, 2009), one can obtain a *reachability matrix* (R) where all (unidirectional) direct *and* indirect relations among the attributes are represented by "1" and the lack of such relation by "0" (Table 2). Note that each attribute is by default related to itself (e.g., A1 is related to A1).

Table 1 *An adjacency matrix (A) based on five attributes*

	A1	A2	A3	A4	A5
A1	0	1	1	0	0
A2	0	0	0	1	0
A3	0	0	0	0	1

A4	0	0	0	0	0
A5	0	0	0	0	0
A5	0	0	0	0	0

Table 2 *A reachability matrix (R) based on five attributes*

	A1	A2	A3	A4	A5
A1	1	1	1	1	1
A2	0	1	0	1	0
A3	0	0	1	0	1
A4	0	0	0	1	0
A5	0	0	0	0	1

The next step involves determining the allowable item types (i.e., potential attribute combinations) based on the specified attributes and their relations. Initially, an *incident matrix* (Q) can be made where the columns represent possible combinations of attributes and the rows represent the specified attributes. In the above example involving five attributes, the number of potential combinations is 31 (that is, 2^5-1) should there be no hierarchical relations among the attributes. However, because of the hierarchy of attributes (Figure 1), not all potential combinations are allowed. For example, an item type that only involves attributes A1 and A4 is not allowed because it is impossible to tap attribute A4 without tapping attribute A2. By removing those unallowable item types, one can obtain a *reduced incident matrix* (Q_r). The reduced Q matrix for our example will look like the following (Table 3), where each column represents one allowed item type and each row represents one attribute.

Table 3 *A reduced incident matrix (Q_r) based on five attributes*

	Item types									
Attributes	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10
A1	1	1	1	1	1	1	1	1	1	1
A2	0	1	0	1	1	1	0	1	1	1
A3	0	0	1	1	0	1	1	0	1	1
A4	0	0	0	0	1	1	0	0	0	1
A5	0	0	0	0	0	0	1	1	1	1

In an ideal scenario where test takers' item responses fully conform to the specified attributes and their hierarchical structure, the 10 item types illustrated in Table 3 can also be seen as test takers' ideal item response patterns. Because the response patterns entail specific combinations of attribute mastery/non-mastery, these patterns represent examinees' various knowledge states. With this understanding, we can construct an *ideal response matrix* (E) where the columns represent different item types and the rows represent test takers' various knowledge states (Table 4). This matrix shows the mappings between test takers' attribute mastery patterns (or knowledge states) and ideal item response patterns (or item types). For example, a test taker

mastering attributes A1, A2, and A3 is expected to respond correctly to item types i1, i2, i3, and i4 (please also refer to the reduced Q matrix in Table 3); however, this test taker is not expected to have correct responses to item type i5, which requires the mastery of attribute A4 for correct response, nor is he/she expected to respond correctly to item type i9 because that requires the mastery of attribute A5 in addition to A1, A2, and A3.

Table 4 *Ideal response matrix (E) based on five attributes*

Attribute mastery patterns (or knowledge states)					Ideal response patterns (item types)									
A1	A2	A3	A4	A5	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10
1	0	0	0	0	1	0	0	0	0	0	0	0	0	0
1	1	0	0	0	1	1	0	0	0	0	0	0	0	0
1	0	1	0	0	1	0	1	0	0	0	0	0	0	0
1	1	1	0	0	1	1	1	1	0	0	0	0	0	0
1	1	0	1	0	1	1	0	0	1	0	0	0	0	0
1	1	1	1	0	1	1	1	1	1	1	0	0	0	0
1	0	1	0	1	1	1	0	0	0	1	0	0	0	0
1	1	0	0	1	1	1	1	1	0	0	1	1	0	0
1	1	1	0	1	1	1	1	1	0	0	1	1	1	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Note. For the "Attribute mastery patterns" section, "1" denotes mastery of an attribute and "0" denotes non-mastery; for the "Ideal response patterns" section, "1" denotes correct response(s) and "0" denotes incorrect response(s)

With the *ideal response matrix* (E), we can infer test takers' latent attribute mastery patterns (i.e., knowledge states) based on their observable item response patterns (i.e., test performance). Note that what is described here assumes an ideal situation where test takers do not produce atypical item responses that do not conform to their attribute mastery patterns (or inconsistent with the attributes that an item is designed to measure). An example of atypical item response is for an examinee mastering only attribute A1 to get correct responses to item type i2. In reality, the ideal situation, as illustrated by the ideal response matrix (E), is virtually impossible to exist, as test takers can always be expected to produce unexpected responses (e.g., a low-ability examinee responds correctly to a high-difficulty item). Hence, there needs to be a means to take into consideration examinees' atypical responses when inferring their latent knowledge states. This brings us to the next phase of the RSM: formulating a classification space.

During the second phase, the formulation of a classification space (or rule space) relies on the calculation of two sets of coordinates: examinees' IRT-based estimation of ability level (or θ) as well as an index indicating how atypical their item response patterns are (or ζ). The classification space can thus be visualized as consisting of a set of ideal points (θ_R, ζ_R) based on the ideal item response patterns, as well as a set of non-ideal points (θ_x, ζ_x) for all test takers based on their actual item response patterns. Each ideal point represents a pre-specified

knowledge state and each non-ideal point represents an examinee's observed pattern of item responses.

In the last phase (i.e., classifying test taker responses into pre-specified knowledge states), Mahalanobis distances (i.e., a statistic used to measure the likelihood ratio between a sample and a population) are calculated between each non-ideal point (θ_x, ζ_x) and each ideal point (θ_R, ζ_R) and the Bayes' classification rule for minimum error is applied to determine which pre-specified knowledge state (represented by the corresponding ideal point) a test taker (represented by the corresponding non-ideal point) belongs to. In this way, individual test takers' mastery and non-mastery of attributes can be diagnosed for subsequent remedial teaching and learning.

Rule Space Method and Its Application to L2 Assessment

The RSM and other diagnostic language assessment methods (e.g., the Fusion Model) have been applied to educational assessment (e.g., math) in order to diagnose learners' mastery of latent cognitive skills and knowledge. In the field of L2 assessment, the application of the RSM and related techniques remains very limited, with a few studies examining L2 learners' knowledge states as they respond to test items assessing reading (Buck, Tatsuoka & Kostin, 1997; Jang, 2009a, 2009b; Kim, 2015), listening (Buck & Tatsuoka, 1998), or both skills (Lee & Sawaki, 2009; Sawaki, Kim, & Gentile, 2009).

In a pioneering study, Buck et al. (1997) applied the RSM to diagnose the sub-skills involved in responding to the reading comprehension section of a TOEIC test among 5,000 Japanese examinees. Based on literature review and test item analyses, the researchers identified 27 potential attributes (e.g., the ability to recognize relevant information, the ability to identify the gist of a passage, the ability to use a word-matching strategy in selecting the correct option, the knowledge of low-frequency vocabulary). Through four rounds of statistical analyses, 24 attributes were retained for examinee classification. Ninety-one percent (91%) of the examinees were successfully classified into one of the knowledge states consisting of the 24 attributes and those attributes together accounted for 97% of the variances in test performance.

Focusing on the reading comprehension sections of two forms of the *LanguEdge* assessment (part of a courseware for preparing the TOEFL iBT), Jang (2009a, 2009b) combined examinee verbal protocol analysis and statistical analysis to identify nine attributes assessed by the reading comprehension test items (e.g., deducing the meaning of a word or a phrase by searching and analyzing a text and by using contextual clues appearing in the text, read carefully or expeditiously to locate relevant information in a text and to determine which information is true or not true). Those nine attributes were used to develop the Q matrix. The *LanguEdge* tests were administered to 2,703 test takers. Different from Buck et al.'s (1997) study, Jang (2009a) applied the Fusion Model for statistical analysis to classify the examinees to three categories (i.e., mastery, non-mastery, and undetermined) for each attribute[2]. The average classification rates were 90% for test Form One and 88% for test Form Two. Jang also reported, among other things, the varying levels of diagnostic capacity of individual test items as well as the usefulness of

diagnostic feedback on improving subsequent teaching. In another study focusing on the TOEFL iBT, Sawaki et al. (2009) relied on expert judgment to identify and code the attributes assessed by two forms of the reading (and listening) comprehension section of the test. A total of 441 examinees completed both forms of the test. After applying the Fusion Model, the researchers finalized a set of four attributes for developing the Q matrix for the reading section (and another four attributes for developing the Q matrix for the listening section). The results showed that, across the two test forms, 76.2% of the examinees were consistently classified into their respective attribute mastery states (for all attributes, or all but one attribute) for the reading section (and 79.6% for the listening section).

In a more recent study focusing on an English placement test, Kim (2015) combined literature search and instructor coding to identify the attributes involved in the reading comprehension section of the test. Ten attributes (e.g., strategy of finding information, strategy of inferencing, knowledge of lexical meaning, knowledge of sentence meaning) were identified for constructing the Q matrix for subsequent statistical analysis. Similar to Jang's (2009a, 2009b) studies cited above, Kim's analysis focused on the mastery probabilities of individual attributes, and reported varied levels of mastery (e.g., ranging from 51.5% to 71.2% across the attributes). The attribute mastery probabilities also differed significantly across beginner, intermediate, and advanced proficiency groups. Finally, the study provided diagnostic reports for individual examinees regarding the degree of mastery of the 10 attributes.

Three observations can be made after summarizing this limited body of empirical research on diagnostic assessment of reading ability. First, although utilizing existing tests may bring concerns of generalizability because researchers need to accommodate the specifics of a particular set of test items in the process of identifying relevant attributes, it remains a common practice in the literature. Second, a related observation is the lack of agreed-upon methods/procedures for identifying attributes. As the above summaries can show, expert judgment, literature search, examinee protocol analysis, and sometimes a combination of these procedures, have been adopted by researchers. The consequence, however, is very different sets of attributes even for the same language skill (e.g., reading) assessed by similar tests (e.g., comparing Jang's (2009a, 2009b) studies and Sawaki et al.'s (2009) study). The question, therefore, is to what extent the identified set of attributes an artifact of the research procedures involved. Because an ultimate goal of diagnostic language assessment is to provide individual examinees with detailed information regarding knowledge/skill mastery for the purpose of remedial learning/instruction, it is important that the attribute mastery reports closely reflect their true ability rather than being influenced by extraneous factors. Finally, previous research has exclusively focused on English as the target language, and it is desirable to extend this line of research to other languages for generalizability considerations. In practice, Chinese is an ideal candidate language, thanks to the growing world-wide popularity of the language. Earlier estimations reported that approximately 30 million people were studying Chinese as a second language around the world (Xu, 2006), and over 3,000 institutions of higher education were

offering Chinese courses (China Educational Newspaper, 2009, September 30). This huge demand in Chinese language learning calls for effective means of assessment that can provide fine-grained information for learners in order to enable sustained learning effort. To this end, this study represents an exploratory effort in the field of L2 Chinese assessment.

Research Question

This study aimed to apply the RSM to analyzing L2 Chinese test takers' responses to the reading comprehension test items of a standardized Chinese proficiency test (i.e., the C. Test). The research question was: Is it feasible to use the RSM to conduct diagnostic assessment of examinees' reading ability in L2 Chinese?

METHOD

Participants

On December 2, 2007, the C. Test (A-D level) was officially administered to 857 test takers globally. All those test takers became our participants. There were 668 Japanese test takers, 139 Koreans, 36 Chinese (ethnic minorities in China with Chinese as their second language), two Filipinos, two Vietnamese, two Malaysians, two Cambodians, two Indians, one Russian, one Australian, one Polish, and one Mauritius. Among these examinees, 681 took the test in Japan, 109 in South Korea, and the remaining 67 in China. The mean test score of the examinee sample was 67.66 (out of 160) and the Standard Deviation (SD) was 27.99. The mean score of the reading comprehension section (detailed below) was 13.51 (out of 30) with an SD of 5.16.

Instrument

The C. Test, or Test of Practical Chinese "实用汉语水平认定考试", is a standardized Chinese proficiency test developed by the Chinese Proficiency Test Center of Beijing Language and Culture University and was launched in 2006. The test has two different proficiency levels, namely, E-F (Elementary) and A-D (Intermediate to Advanced)[3]. The instrument used in this study was the reading comprehension section of the C. Test (A-D) officially administered on December 2, 2007. In this version of the test, there were six reading comprehension texts each with five multiple-choice questions (each contained four options) for a total of 30 items. The texts were 714 to 803 characters in length and the content did not require specialized knowledge. Readers interested in accessing the test items can refer to HSK Center (2008).

Procedures

Attribute identification involved several procedures. The researchers first consulted published empirical research on diagnostic assessment of L2 reading ability and theories of reading comprehension to prepare a list of potentially relevant attributes. This list and the 30 test items were then forwarded to two domain experts in L2 Chinese reading comprehension, who

identified nine attributes that were assessed by the test items. Afterwards, the researchers recruited 10 Chinese language instructors with minimally five years of teaching experience to review and code the test items according to the nine attributes. Following Kim (2015), an attribute with over 60% agreement among the coders for each test item was considered essential and subsequently retained. As it turned out, one attribute, the ability to apply background knowledge, was measured by less than three items. Following Kim (2015), this attribute was removed from the original attribute list. Table 5 shows the remaining eight attributes with their corresponding item characteristics. Finally, the eight attributes and the item codings were reviewed by the two domain experts, who discussed and finalized the attribute hierarchy illustrated in Figure 2.

Table 5 *Attribute list for the C. Test reading comprehension section*

Attribute	Item characteristics coded
A1. Ability to recognize characters and words	Correct response to the item entails appropriate knowledge of Chinese characters and words
A2. Ability to hold information in memory	The options tend to be long, and/or the necessary information spreads over two sentences
A3. Ability to use given information as a basis for searching the text	The necessary information or information in options is easy to locate
A4. Ability to understand explicitly stated information	The item requires understanding of literal meaning of words and sentences
A5. Ability to understand the gist of a passage	The item is a "main idea" item
A6. Ability to recognize relevant information	The necessary information occurs out of item order, and/or the necessary information is scattered across the text
A7. Ability to understand implicit/implied meaning and/or attitude	The necessary information (e.g., meaning and/or attitude) is not explicitly stated and needs to be inferred
A8. Ability to infer word meaning in context	The item asks for the meaning of a specific word and/or phrase appeared in the text

Following the procedures outlined in the literature review section, we constructed the *adjacency matrix* (A), the *reachability matrix* (R), the *incident matrix* (Q), the *reduced incident matrix* (Q_r), and the *ideal response pattern* (E). Because there were eight attributes involved, the

incident matrix (Q) included 2^8-1=255 possible combinations of attributes; however, because of the hierarchical structure among the attributes (Figure 2), only 52 of the 255 combinations were allowed. These 52 combinations were included in the *reduced incident matrix* (Q_r) shown in Table 6. Then, a 52 x 52 *ideal response pattern* (E) was developed with the rows representing possible knowledge states of examinees and the columns representing different item types ~~(Appendix A)~~.

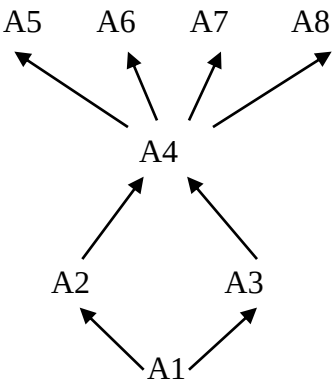

Figure 2. Hierarchical structure of eight attributes assessed by the reading comprehension section of the C. Test.

Table 6 *The reduced incident matrix (Q_r) based on the eight attributes*

	Item type												
Attribute	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10	i11	i12	i13
A1	1	1	1	1	1	1	1	1	1	1	1	1	1
A2	0	1	1	1	1	1	1	1	1	1	1	1	1
A3	0	0	0	0	0	0	0	0	0	0	0	0	0
A4	0	0	1	1	1	1	1	1	1	1	1	1	1
A5	0	0	0	1	0	0	0	1	1	1	0	0	0
A6	0	0	0	0	1	0	0	1	0	0	1	1	0
A7	0	0	0	0	0	1	0	0	1	0	1	0	1
A8	0	0	0	0	0	0	1	0	0	1	0	1	1

	Item type												
Attribute	i14	i15	i16	i17	i18	i19	i20	i21	i22	i23	i24	i25	i26
A1	1	1	1	1	1	1	1	1	1	1	1	1	1
A2	1	1	1	1	1	0	0	0	0	0	0	0	0
A3	0	0	0	0	0	1	1	1	1	1	1	1	1
A4	1	1	1	1	1	0	1	1	1	1	1	1	1
A5	1	1	0	0	1	0	0	1	0	0	0	1	1
A6	1	0	1	1	1	0	0	0	1	0	0	1	0
A7	1	1	0	1	1	0	0	0	0	1	0	0	1
A8	0	1	1	1	1	0	0	0	0	0	1	0	0

	Item type												
Attribute	i27	i28	i29	i30	i31	i32	i33	i34	i35	i36	i37	i38	i39
A1	1	1	1	1	1	1	1	1	1	1	1	1	1
A2	0	0	0	0	0	0	0	0	0	1	1	1	1
A3	1	1	1	1	1	1	1	1	1	1	1	1	1
A4	1	1	1	1	1	1	1	1	0	1	1	1	1
A5	1	0	0	0	1	1	1	0	1	0	0	1	0
A6	0	1	1	0	1	0	1	1	1	0	0	0	1
A7	0	1	0	1	1	1	0	1	1	0	0	0	0
A8	1	0	1	1	0	1	1	1	1	0	0	0	0

	Item type												
Attribute	i40	i41	i42	i43	i44	i45	i46	i47	i48	i49	i50	i51	i52
A1	1	1	1	1	1	1	1	1	1	1	1	1	1
A2	1	1	1	1	1	1	1	1	1	1	1	1	1
A3	1	1	1	1	1	1	1	1	1	1	1	1	1
A4	1	1	1	1	1	1	1	1	1	1	1	1	1
A5	0	0	1	1	1	0	0	0	1	1	1	0	1
A6	0	0	1	0	0	1	1	0	1	1	0	1	1
A7	1	0	0	1	0	1	0	1	1	0	1	1	1
A8	0	1	0	0	1	0	1	1	0	1	1	1	1

The next step was to calculate a set of coordinates consisting of examinees' IRT-based estimation of ability level (or θ) and their atypical response index (or ζ). Because IRT-based parameter estimation cannot be made for examinees who answer all items correctly or incorrectly as well as for items that all examinees answer correctly or incorrectly (Hambleton, Swaminathan & Rogers, 1991), the first and last rows and the first and last columns of the *ideal response pattern* (E) were removed, resulting in 50 rows and 50 columns in the ideal response pattern (E) for subsequent statistical analyses. We calculated 50 ideal points (θ_R, ζ_R) based on the ideal item response patterns as well as 857 non-ideal points (θ_x, ζ_x) based on the examinees' actual item response patterns.

Finally, in order to classify the examinees into the 50 pre-specified knowledge states, we calculated Mahalanobis distances (D^2) between each non-ideal point (θ_x, ζ_x) and each ideal point (θ_R, ζ_R). Because Mahalanobis distances (D^2) follow the X^2 distribution with two degrees of freedom (Tatsuoka & Tatsuoka, 1987), D^2 less than 5.99 is considered to be valid for classification. For an examinee who met this criteria (i.e., D^2 less than 5.99), he/she was classified into the nearest pre-specified knowledge state based on the smallest D^2.

RESULTS

The results showed that 776 of the 857 test takers' Mahalanobis distances (D^2) were smaller than 5.99, and they were subsequently classified into 39 of the 50 pre-specified knowledge states. The

classification rate was thus 90.54%. Table 7 shows the 50 pre-specified knowledge states (i.e., attribute mastery patterns, where "1" stands for mastery and "0" stands for non-mastery), their corresponding coordinates (θ_R, ζ_R) based on the ideal response pattern (E), and the number and percentage of participants (*N*=857) classified into each knowledge state. In reviewing Table 7, it is clear that the majority of the examinees (67.1%) were found to belong to nine knowledge states, namely, #37 (17.9%, mastery of A1, A2, A3, A4), #32 (9.5%, mastery of A1, A3, A4, A5, A7, A8), #38 (7.7%, mastery of A1, A2, A3, A4, A5), #30 (7.0%, mastery of A1, A3, A4, A7, A8), #36 (5.4%, mastery of A1, A2, A3), #51 (5.4%, mastery of A1, A2, A3, A4, A6, A7, A8), #34 (4.9%, mastery of A1, A3, A4, A6, A7, A8), #24 (4.7%, mastery of A1, A3, A4, A8), #12 (4.6%, mastery of A1, A2, A4, A6, A8) (refer to Table 5 for details of attributes A1-A8). However, there was no predominant knowledge state(s): even though knowledge state #37 represented the profiles of the largest group of examines (*N*=154), the percentage score showed that it was still a relatively small portion of the examinee sample (i.e., 17.9%).

Table 7 *Classification results based on 50 pre-specified knowledge states*

Number	Attribute mastery patterns (A1, A2, ..A8)	(θ_R, ζ_R)	Number of participants classified (percentage)
1	10000000	N/A*	N/A*
2	11000000	(-1.6054, 0.1621)	6 (0.7%)
3	11010000	(-1.2695, -0.3274)	2 (0.2%)
4	11011000	(-0.9928, -0.1850)	12 (1.4%)
5	11010100	(-1.0212, -0.2519)	0 (0.0%)
6	11010010	(-1.0138, -0.2341)	0 (0.0%)
7	11010001	(-1.0204, -0.2499)	0 (0.0%)
8	11011100	(-0.5535, 0.1307)	1 (0.1%)
9	11011010	(-0.5410, 0.1540)	1 (0.1%)
10	11011001	(-0.5530, 0.1320)	2 (0.2%)
11	11010110	(-0.5622, 0.1067)	2 (0.2%)
12	11010101	(-0.6530, -0.0270)	40 (4.6%)
13	11010011	(-0.5615, 0.1082)	0 (0.0%)
14	11011110	(0.0578, 0.9069)	0 (0.0%)
15	11011011	(0.0582, 0.9075)	9 (1.0%)
16	11011101	(-0.0035, 0.8112)	9 (1.0%)
17	11010111	(0.0033, 0.8277)	6 (0.7%)
18	11011111	(0.6744, 3.3664)	3 (0.3%)
19	10100000	(-1.6091, 0.1455)	2 (0.2%)
20	10110000	(-1.2758, -0.3491)	10 (1.1%)
21	10111000	(-1.0051, -0.2163)	1 (0.1%)
22	10110100	(-1.0205, -0.2528)	1 (0.1%)
23	10110010	(-1.0230, -0.2588)	0 (0.0%)
24	10110001	(-1.0245, -0.2624)	41 (4.7%)
25	10111100	(-0.5579, 0.1227)	0 (0.0%)
26	10111010	(-0.5598, 0.1185)	0 (0.0%)
27	10111001	(-0.5679, 0.1047)	2 (0.2%)

28	10110110	(-0.5642, 0.1039)	0 (0.0%)
29	10110101	(-0.5653, 0.1014)	12 (1.4%)
30	10110011	(-0.5245, 0.1764)	60 (7.0%)
31	10111110	(0.0500, 0.8957)	0 (0.0%)
32	10111011	(-0.0930, 0.6170)	82 (9.5%)
33	10111101	(0.0374, 0.8717)	4 (0.4%)
34	10110111	(0.0724, 0.9823)	42 (4.9%)
35	10111111	(0.7015, 3.4383)	1 (0.1%)
36	11100000	(-0.7387, 0.1336)	47 (5.4%)
37	11110000	(0.0268, -0.9400)	154 (17.9%)
38	11111000	(0.5322, -1.1810)	66 (7.7%)
39	11110100	(0.3088, -1.5182)	0 (0.0%)
40	11110010	(0.3103, -1.5202)	6 (0.7%)
41	11110001	(0.3080, -1.5171)	13 (1.5%)
42	11111100	(0.9076, -1.4942)	3 (0.3%)
43	11111010	(0.9120, -1.5185)	15 (1.7%)
44	11111001	(0.9052, -1.4847)	26 (3.0%)
45	11110110	(0.7469, -1.7942)	7 (0.8%)
46	11110101	(0.8009 , -1.5488)	10 (1.1%)
47	11110011	(0.7630, -1.6808)	1 (0.1%)
48	11111110	(1.5989, -0.8373)	10 (1.1%)
49	11111101	(1.5267, -0.6879)	13 (1.5%)
50	11111011	(1.6082, -0.7634)	7 (0.8%)
51	11110111	(1.4378, -0.0598)	47 (5.4%)
52	11111111	N/A*	N/A*

Note. * These two knowledge states (#1, #52) were removed from final analysis, as discussed earlier.

With a classification rate of 90.54%, it means that 9.45% (or 81) examinees were not successfully classified. As it turned out, these unclassified examinees tended to have either relatively higher or relatively lower ability: among the 81 examinees, eighteen (or 22.22%) fell out of ± 2 SDs and 57 (70.37%) fell outside ± 1 SD along the ability axle. Moreover, the percentage of unclassified examinees tended to be higher among below-average-ability examinees (i.e., whose *z* scores were below zero) than among above-average-ability ones (i.e., whose *z* scores were above zero). In this study, there were 463 below-average-ability examinees, among which 54 (or 11.66%) were unclassified. In contrast, among the 394 above-average-ability examinees, 27 (or 6.85%) were unclassified. In other words, below-average-ability examinees were nearly twice as likely to be unclassified as above-average-ability examinees.

Table 8 further shows the mastery levels of the eight attributes for the entire examinee group. As expected, the level of mastery varied considerably across the eight attributes, with A1 (The ability to recognize characters and words) being the best mastered skills and A6 (The ability to recognize relevant information) being the least mastered skill. In general, the attributes located at the lower part of the hierarchy were better mastered than those located at the upper part of the

hierarchy (refer to Figure 2 for details on the hierarchy).

Table 8 *Percentage of attribute mastery*

Attribute	Percentage of mastery
A1. Ability to recognize characters and words	100.0%
A2. Ability to hold information in memory	60.4%
A3. Ability to use given information as a basis for searching the text	79.6%
A4. Ability to understand explicitly stated information	84.1%
A5. Ability to understand the gist of a passage	31.1%
A6. Ability to recognize relevant information	24.6%
A7. Ability to understand implicit/implied meaning and/or attitude	34.8%
A8. Ability to infer word meaning in context	44.6%

Because one advantage of diagnostic language assessment is to provide detailed information about individual test takers' strengths and weaknesses of targeted linguistic domain, we are able to provide individualized diagnostic reports for those successfully classified examinees. Due to space limit, we juxtaposed two such reports for two examinees in Table 9. It is interesting that, although the two examinees were at the same overall ability level ($\theta = 0.2029$), their knowledge patterns differed. Examinee 1 was classified into knowledge state #34, meaning that he/she had already mastered attributes A1 (ability to recognize characters and words), A3 (ability to use given information as a basis for searching the text), A4 (ability to understand explicitly stated information), A6 (ability to recognize relevant information), A7 (ability to understand implicit/implied meaning and/or attitude), and A8 (ability to infer word meaning in context), and that he/she had yet to master attributes A2 (ability to hold information in memory) and A5 (ability to understand the gist of a passage). In contrast, Examinee 2 was classified into knowledge state #37, which means that he/she had mastered attributes A1, A2, A3, and A4, but not attributes A5, A6, A7, and A8.

Finally, in reviewing our data, we also found that test takers with different ability levels belonged to the same knowledge states. For example, we found two examinees with their respective ability levels (θ) of -0.4879 and -0.1562, yet the classification results showed that they both belonged to knowledge state #37, meaning that they both mastered attributes A1, A2, A3 and A4, but not attributes A5, A6, A7, and A8.

Table 9 *Two sample diagnostic reports of reading comprehension ability*

Examinee	Examinee 1	Examinee 2
Examinee ability level	0.2029	0.2029
Attribute mastery pattern	#34 (10110111)	#37 (11110000)

Attributes already mastered	A1. Ability to recognize characters and words A3. Ability to use given information as a basis for searching the text A4. Ability to understand explicitly stated information A6. Ability to recognize relevant information A7. Ability to understand implicit/implied meaning and/or attitude A8. Ability to infer word meaning in context	A1. Ability to recognize characters and words A2. Ability to hold information in memory A3. Ability to use given information as a basis for searching the text A4. Ability to understand explicitly stated information
Attributes to be mastered	A2. Ability to hold information in memory A5. Ability to understand the gist of a passage	A5. Ability to understand the gist of a passage A6. Ability to recognize relevant information A7. Ability to understand implicit/implied meaning and/or attitude A8. Ability to infer word meaning in context

DISCUSSION

The purpose of this study was to explore the feasibility of applying the RSM to diagnostic assessment of reading comprehension ability among learners of L2 Chinese. The results showed that 90.54% of the 857 examinees who took the test were successfully classified into the pre-specified knowledge states. This classification rate was comparable with those reported by Buck et al. (1997) and by Jang (2009a, 2009b) and was higher than those reported by Sawaki et al. (2009). According to Tatsuoka (2009), a classification rate of 90% is an important indicator of the validity of a RSM study in showing that the proposed attributes and their relationship as illustrated in the Q matrix (Table 6) fit our examinees' performance well.

Nevertheless, 9.46% of our examinees were not successfully classified. A closer examination suggested that these examinees did not seem to follow a pattern of normal distribution in their overall level of reading abilities; rather, they were much more likely to have either relatively higher- or relatively lower- ability (i.e., outside the range of ± 1 SD). Moreover, below-average-ability examinees appeared to be more likely to be unclassified than their above-average-ability counterparts. While the exact reason for these observations could not be identified based on the data we have collected, one possibility, as also expressed by Buck & Tatsuoka (1998), is that certain attribute(s) that influenced those examinees' test performance

remained unidentified, which, in turn, means that certain knowledge state(s) that could explain those examinees' performance were not included in our analysis reported earlier. Lower-ability or higher-ability examinees might particularly utilize certain types of knowledge and/or cognitive skill(s) in responding to test items. However, without probing those examinees' online processing procedures involved in completing the test, it would be difficult to identify such knowledge and/or skill. Future research will need to employ techniques, such as a think-aloud protocol or stimulated recall, to assist with attribute identification.

Our results also showed that the examinees' knowledge states were highly diverse, covering 39 of the 50 pre-specified knowledge states. This diverse distribution of knowledge states provides a refined illustration of the individual differences in reading comprehension ability among the examinees. In this study, reading comprehension ability, as measured by 30 reading comprehension test items, was indexed through the mastery and non-mastery of eight attributes. Because each knowledge state represented a specific combination of mastered and un-mastered attributes, our results showed the details of 39 types of reading comprehension ability profiles among the successfully classified examinees (see Table 7). In this way, a test score (or rather, a test result) becomes readily interpretable in terms of the strengths and weakness of the targeted domain of linguistic competence (i.e., reading comprehension ability).

The ease of test score (or result) interpretation, as afforded by diagnostic language assessment, can effectively facilitate the development of on-target remedial instruction and learning activities by pointing out the specific learning objectives. This point is illustrated in two scenarios extracted from this study. In the first scenario, regardless of their overall ability level, the examinees classified into the same knowledge state would benefit from the same instructional/learning package aiming at developing those yet-to-be-mastered attributes. In the second scenario, examinees with the same overall ability level might actually need different instructional/learning packages due to variations in attribute mastery patterns. The two examinee profiles illustrated in Table 9 are a good example here: despite their identical overall ability level, Examinee 1 belonged to knowledge state #34 while Examinee 2 was classified into knowledge state #37. Together, these two scenarios showed the risks of relying on a single holistic ability measure in guiding the development of remedial instruction, and pointed to the advantage of diagnostic language assessment in providing refined objectives for subsequent instruction and learning. Pedagogically, the implication is that, for the purpose of developing complex language skills such as reading comprehension that consist of multiple attributes, an effective instructional program should be desgined at the level of attributes in order to allow individualized remedial teaching and learning.

Finally, at the level of individual attributes, our results were consistent with previous studies (e.g., Buck et al. 1997; Jang, 2009b; Kim, 2015) in showing that the degree of mastery of individual attributes varied considerably, ranging from 24.6% for A6 to 100.0% for A1. In reviewing Table 9 along with Figure 2 that illustrates the hierarchical structure of the attributes, it becomes clear that the attributes located at the lower portion of the hierarchy were better

mastered than those located in the higher portion of the hierarchy. The only exception is A4, which, although located in a higher position in the hierarchy than A2 and A3, showed a better degree of mastery than the other two. This result can be explained by the structure of hierarchy, that is, there are two routes toward mastering A4, one through the mastery of A1 and A2, and the other one through the mastery of A1 and A3. In other words, in addition to the mastery of A1, mastering either A2 or A3 constitutes a necessary, but not sufficient, condition for the mastery of A4; hence, the finding that A4 exhibited a higher level of mastery than A2 and A3 is not unexpected. Overall, the mastery levels across the eight attributes as shown in Table 8 can lend support to the validity of the proposed attribute hierarchy – it makes good sense that more basic skills are mastered before more advanced skills.

LIMITATIONS AND ISSUES FOR FUTURE EXPLORATION

This study explored the feasibility of applying the RSM to diagnostic assessment of reading comprehension ability in L2 Chinese. The findings suggest that the RSM can be a useful technique for providing the majority of the examinees (over 90%) with fine-grained information about their mastery and non-mastery of attributes assessed by a reading comprehension test. However, as this study represented an initial effort in diagnostic assessment for L2 Chinese, it was limited in several ways, and future studies are needed to refine this line of research.

To begin with, although the classification rate was above 90% and can thus be considered as successful from a research point of view, it is also true that nearly 10% of the examinees were not classified. If diagnostic language assessment is to be put into practice, we cannot afford to provide diagnostic information only to a subset of examinees. In fact, no previous study utilizing existing tests has achieved a classification rate of 100%. This means researchers will have to examine what factors contribute to unsuccessful classification. In this study, unsuccessful classification occurred when an examinee's test response pattern could not be categorized into any pre-specified knowledge states with an acceptable level of confidence (p <.05). As mentioned above, this was most likely due to incomplete extraction of attribute(s) (and, in turn, knowledge states) for examinees with relatively higher- and relatively lower- levels of ability. Conducting focused investigations into those examinees' cognition involved in reading comprehension, combined with multiple procedures for identifying and selecting attributes (as illustrated in Jang's (2009b) study), seems to be a potential solution. The problem, however, is that those post hoc procedures are inevitably influenced by the characteristics of specific test items as well as the theories and empirical findings that researchers consult with, and this is perhaps why researchers have had different sets of attributes for the same language skill assessed by similar tests. In this regard, this study was limited in that the analyses were based on an existing test (i.e., C. Test) and therefore encountered the same issues reported in previous studies (e.g., Buck et al. 1997; Jang, 2009a, 2009b).

An alternative, and probably better, solution to the above issues is to design and develop

tests for the purpose of diagnostic assessment. This would involve specifying key attributes in the first place and developing test items accordingly. In this way, the influence of extraneous variables (e.g., attributes assessed by a very small number of test items) on examinee classification could be reduced. So far, little empirical effort has been made to examine the feasibility of this approach, and future research is in order.

In terms of participant sampling, the fact that our test takers were predominantly Asian tends to constrain the generalizability of the findings to all L2 Chinese learners. In this study, we made an effort to include all official test takers of a particular test administration, thus our sample, in a realistic sense, did reflect the examinee population of the test. However, it is interesting to note that our examinees were classified into 39 of the 50 pre-specified knowledge states. While the variety of the knowledge states found among our examinees could be counted as evidence to support an argument that the findings are generalizable to a larger examinee population, whether the remaining 11 pre-specified knowledge states are more likely to be found among non-Asian learners of L2 Chinese would be an interesting question to explore in the future. Likewise, whether the overall classification rate as well as the (major) patterns of reading mastery would remain comparable for non-Asian examinees also awaits future research. Another interesting research topic is to examine whether there is any difference between heritage and non-heritage learners, given the differences in learning opportunities afforded by their respective learning environments.

Finally, because an important goal of diagnostic language assessment is to provide guidance for subsequent remedial teaching and learning, it is necessary to conduct follow-up studies to examine the usefulness of diagnostic information. With few exceptions (e.g., Jang, 2009a), the field has yet to pay sufficient attention to this area.

REEFERENCES

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedures to language testing: Examining attributes of a free response listening test. *Language Testing*, *15*(2), 119-157.

Buck, G., Tatsuoka, K., & Kostin, I. (1997). The sub-skills of reading: Rule-space analysis of a multiple-choice test of second language reading comprehension. *Language Learning*, *47*(3), 423-466.

China Educational Newspaper. (2009, September 30). International promotion of Chinese language and culture, p.4. (中国教育报. (2009 年 9 月 30 日). 中国文化大规模走出国门。第 4 版。).

Gierl, M. J. (2007). Making diagnostic inferences about cognitive attributes using the Rule-Space Model and attribute hierarchy method. *Journal of Educational Measurement*, *44*, 325-340.

Hambleton, R., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. SAGA publications.

HSK Center (2008). *A collection of authentic C-Test administered in 2007*. Beijing: Beijing Language and Culture University Press. (北京语言大学汉语水平考试中心(2008).

Li, S., & Wang J. (2017). Diagnostic assessment of L2 Chinese learners' reading comprehension ability. In Zhang, D. & Lin, C. (Eds.). *Chinese as a second language assessment* (pp. 183–202). Springer.

《C.Test 实用汉语水平认定考试 2008 年真题集》. 北京: 北京语言大学出版社.)

Jang, E. E. (2009a). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing*, *26*(1), 31-73.

Jang, E. E. (2009b). Demystifying a Q-matrix for making diagnostic inferences about L2 reading skills. *Language Assessment Quarterly*, *6*, 210-238.

Kim, A-Y. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing*, *32*(2), 227-258.

Lee, Y-W. (2015). Diagnosing diagnostic language assessment. *Language Testing, Special Issue*, 1-18.

Lee, Y-W., & Sawaki, Y. (2009). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Language Assessment Quarterly*, *6*, 239-263.

Sawaki, Y., Kim, H-J., & Gentile, C. (2009). Q-Matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Language Assessment Quarterly*, *6*, 190-209.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*(4), 345-354.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichos, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327-359). Hillsdale, NJ: Erlbaum.

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. Routledge.

Tatsuoka, K. K., & Tatsuoka, M. M. (1987). Bug distribution and pattern classification. *Psychometrika*, *52*(2), 193-206.

Xu, L. (2006). It is a great good thing to accelerate the international promotion of Chinese language. *Applied Linguistics*, *6*, 8-12. (许琳 (2006). 汉语加快走向世界是件大好事。《语言文字应用》，2006 年 6 月，8-12 页。)

NOTES

1. As Sawaki, Kim & Gentile (2009) summarized, there are three approaches to identifying attributes: (a) by examining surface test item characteristics, (b) by referring to theoretical taxonomies of language ability, and (3) by analyzing test takers' reported skills and processes. The current study adopted the first approach because it was based on an existing test.

2. The Fusion Model and the RSM are similar in that they are both probabilistic models that decompose examinee abilities into cognitive attributes based on a Q Matrix. They are different in terms of assumptions of attribute structure, flexibility of handling items scored polytomously, and parameter estimation methods. Interested readers can refer to Lee & Sawaki (2009) for

comparing the features of various cognitive diagnostic models.

3. The C-Test (A-D) includes two main components: listening comprehension (70 items) and integrated skill use (90 items). The listening comprehension component further includes four sections: (a) graph-based listening comprehension (10 items), (b) short-dialogue-based listening comprehension (20 items), (c) extended-dialogue-based listening comprehension (10 items), and (d) listening comprehension and note-taking (20 items). The integrated skill use component includes six sections: (1) vocabulary/structure (10 items), (2) word order (20 items), (3) reading comprehension (30 items), (4) error identification (10 items), (5) passage-based blank filling (10 items), and (6) passage-based sentence making (10 items). The allowable time for completing the entire test is 150 minutes (i.e., 50 for listening and 100 for integrated skill use).

~~APPENDIX A~~

~~Ideal response pattern (E) for the C-Test reading comprehension section.~~

Id	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11	I12	I13	I14	I15	I16	I17	I18	I19	I20	I21	I22	I23	I24	I25	I26
0001	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0002	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0003	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0004	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0005	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0006	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0007	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0008	1	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0009	1	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0010	1	1	1	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0011	1	1	1	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0012	1	1	1	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0013	1	1	1	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0014	1	1	1	1	1	1	0	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
0015	1	1	1	1	0	1	1	0	1	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
0016	1	1	1	1	1	0	1	1	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
0017	1	1	1	0	1	1	1	0	0	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0
0018	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
0019	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
0020	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
0021	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0
0022	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0
0023	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0
0024	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0
0025	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	1	0
0026	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	1
0027	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	1	0	0
0028	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	0	0	0
0029	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	1	0	0
0030	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0	0
0031	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	1	1
0032	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1	0	1
0033	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	1	1	0
0034	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1	0	0
0035	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1

0036	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
0037	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
0038	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0
0039	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0
0040	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0
0041	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0
0042	1	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	1	0
0043	1	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	1
0044	1	1	1	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1	1	1	0	0	1	0	0
0045	1	1	1	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	1	1	0	0	0
0046	1	1	1	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	1	1	0	1	0	1	0	0
0047	1	1	1	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	1	1	0	0
0048	1	1	1	1	1	1	0	1	1	0	1	0	0	1	0	0	0	0	1	1	1	1	1	0	1	1
0049	1	1	1	1	1	0	1	1	0	1	0	0	0	0	0	1	0	0	1	1	1	1	0	1	1	0
0050	1	1	1	1	0	1	1	0	1	1	0	0	1	0	1	0	0	0	1	1	1	0	1	1	0	1
0051	1	1	1	0	1	1	1	0	0	0	1	1	1	0	0	0	1	0	1	1	0	1	1	1	0	0
0052	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Id	I27	I28	I29	I30	I31	I32	I33	I34	I35	I36	I37	I38	I39	I40	I41	I42	I43	I44	I45	I46	I47	I48	I49	I50	I51	I52
0001	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0002	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0003	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0004	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0005	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0008	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0009	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0010	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0011	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0012	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0013	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0014	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0016	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0017	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0018	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0019	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0020	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0021	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0022	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0023	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0024	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0025	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0026	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0027	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0028	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0029	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0030	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0031	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0032	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0033	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0034	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0035	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Li, S., & Wang J. (2017). Diagnostic assessment of L2 Chinese learners' reading comprehension ability. In Zhang, D. & Lin, C. (Eds.). *Chinese as a second language assessment* (pp. 183–202). Springer.

0036	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0037	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0038	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0039	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
0040	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
0041	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
0042	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0
0043	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0
0044	1	0	0	0	0	0	0	0	0	1	1	1	0	0	1	0	0	1	0	0	0	0	0	0	0
0045	0	1	0	0	0	0	0	0	0	1	1	0	1	1	0	0	0	0	1	0	0	0	0	0	0
0046	0	0	1	0	0	0	0	0	0	1	1	0	1	0	1	0	0	0	1	1	0	0	0	0	0
0047	0	0	0	1	0	0	0	0	0	1	1	0	0	1	1	0	0	0	0	0	1	0	0	0	0
0048	0	1	0	0	1	0	0	0	0	1	1	1	1	1	0	1	1	0	1	0	0	1	0	0	0
0049	1	0	1	0	0	0	1	0	0	1	1	1	1	0	1	1	0	1	0	1	0	0	1	0	0
0050	1	0	0	1	0	1	0	0	0	1	1	1	0	1	1	0	1	1	0	0	1	0	0	1	0
0051	0	1	1	1	0	0	0	1	0	1	1	0	1	1	1	0	0	0	1	1	1	0	0	0	1
0052	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1